Note: any in-text citations denoted with a * point readers to more information on the method described

Paul Merica

Mark Fredrickson

5/1/2019

Introduction

In 2008, the most significant recession since the Great Depression hit the United States, and later the entire world. Many areas are still feeling the effects of this recession today. It has been a prevailing economic theory that people have been flocking to bigger cities to look for jobs(Rafter rejournals.com). Many have thought this has appeared to be accelerated by the recession and job scarcity. A simple conclusion from this would be that states with big cities would recover more completely from the recession of 2008 than states that did not have big cities. The point of this research paper will be to investigate whether states with bigger cities(as defined in this paper as having a population of >500,000) have recovered differently in terms of unemployment than states without big cities.

Unemployment is one of the most used indicators of economic health for states, and one that is argued about constantly. In politics, discouraged workers are often used to demean, and devalue high or low unemployment numbers. Discouraged workers are workers that would work if given the opportunity, but have decided not to pursue job searches due to previous failures. Donald Trump criticized Obama by saying that although he had low unemployment numbers, there was a large number of discouraged workers that had not recovered from the recession(Horsley www.npr.org). This isn't a convincing argument to most economists, because the employment bureau hasn't changed the way it calculates unemployment for a long time. Despite this, many voters still believe in this argument, and use in political conversations every day. This is why I will analyze the number of discouraged workers, and underemployed workers along with the tradition rate of unemployment to test the economic health of states.

The point of my research will be to determine whether states with more metropolitan cities(defined in my research as cities with a population greater than 500,000) recovered differently in terms of unemployment than states without big cities. There is a theory that jobs are all in these metropolitan areas in current times, so the rural areas of America are continuing to get less populated, and thus less economic activity is occurring in those rural areas. The continuing departure from rural areas should increase unemployment in these states with no big cities and mean that these rural areas are getting hit the worst by the effects of the recession.

Inversely, states with these metropolitan areas should be experiencing more economic growth and thus recovering faster from the recession. They should have lower unemployment rates, and in hopefully lower than pre-recession levels. My project will look at whether these states recovered more thoroughly than their counterparts, and whether there is a significant

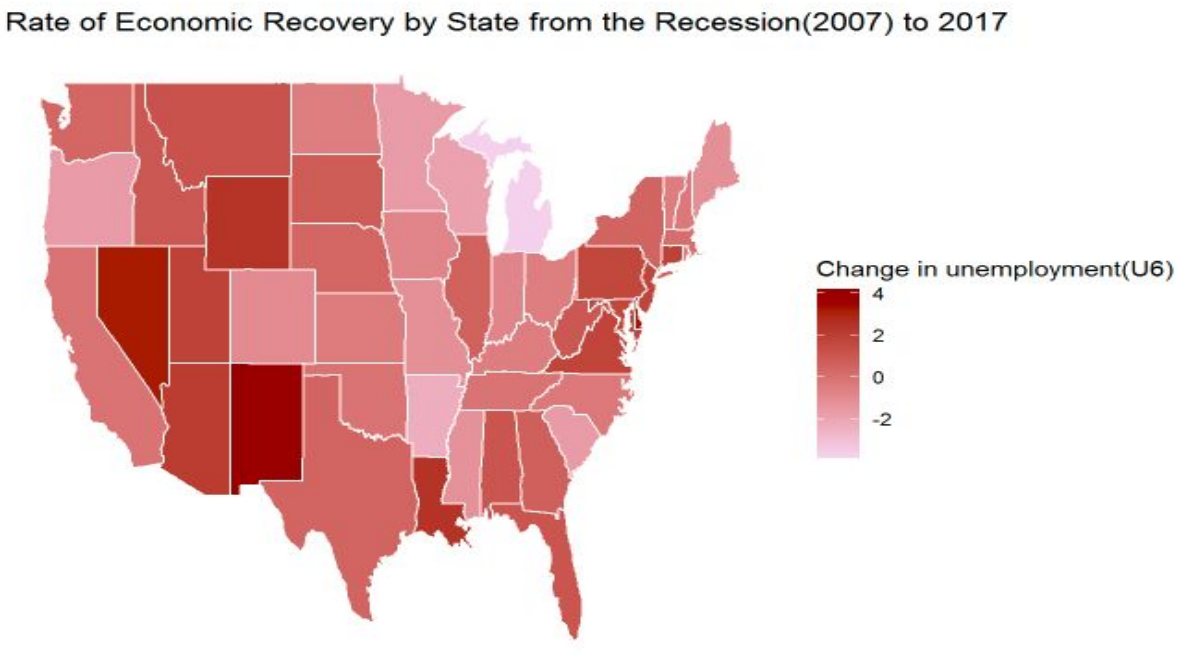
Note: any in-text citations denoted with a * point readers to more information on the method described

difference between the two population means of change in unemployment from 2007 to 2017. If there is a significant difference, I will move on, and see if containing more metropolitan cities in a state is an acceptable estimator from the recession.

Data

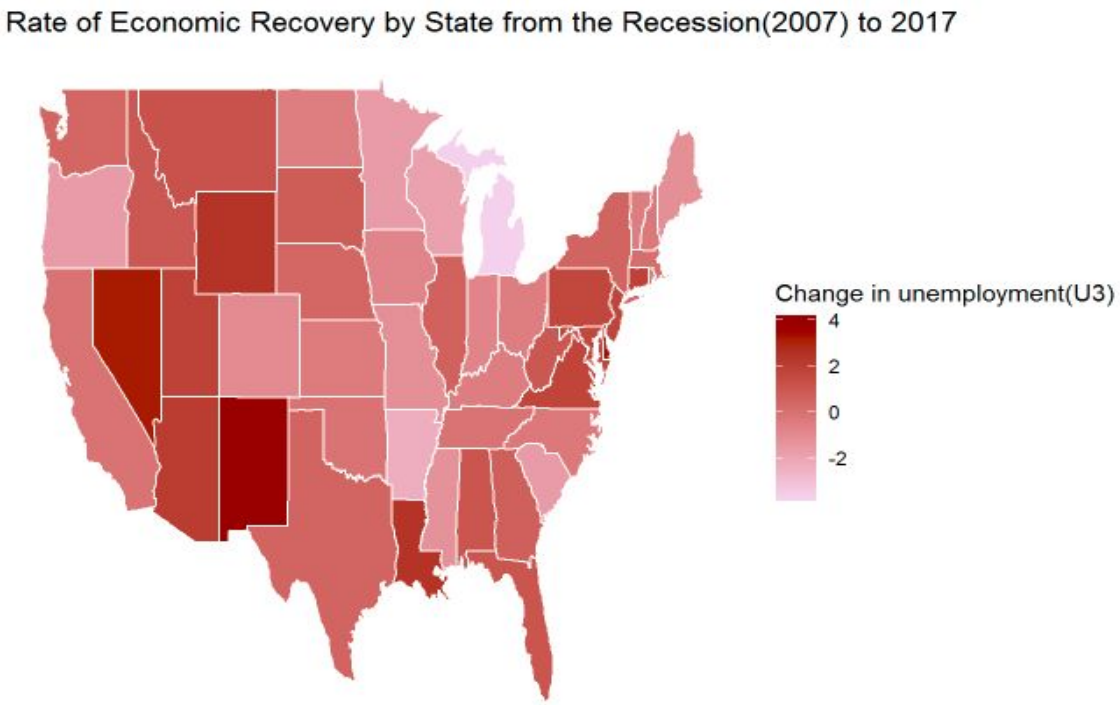
The first dataset I have collected is from the Bureau of Labor Statistics(BLS, www.bls.gov). The data is sorted by state and year. The first dataset I have labeled in my supplemental code as "myprojectdataset". This dataset contains the unemployment rates from 2003 to 2018 for each State. The 2018 data is incomplete though, so we will be using 2017 as the most recent unemployment rate year for this experiment. The dataset has unemployment sorted into six categories U1-U6.

The U3 unemployment rate is the traditional measure used by most economists, as it only includes people in the labor force looking for jobs. For this project, I will be using the U6 unemployment rate. The U6 unemployment rate includes discouraged workers, all other marginally attached workers, and those who are part-time purely for economic reasons. The last category(those employed part-time for economic purposes) are usually referred to as underemployed peoples. They would prefer a full-time job, but haven't found a good match, so they are working part-time to pay the bills. These are people who aren't using their skills to the best benefit of the economy. I sorted by each state and calculated the change by subtracting the 2007 unemployment rate from the 2018 unemployment rate. Doing this allowed me to create this chart of the United States, showing each state's change in unemployment(U6) rate from 2007 to 2017. I also used the code provided by a user on R-Bloggers to get the graphical representation of the United States by state(www.r-bloggers.com).



Note: any in-text citations denoted with a * point readers to more information on the method described

This allows us to see how each state's U6 unemployment rate has recovered from the recession until now. To see if there were any differences if I used U3 instead of U6, I made the same plot but instead using the U3 values provided by the census.



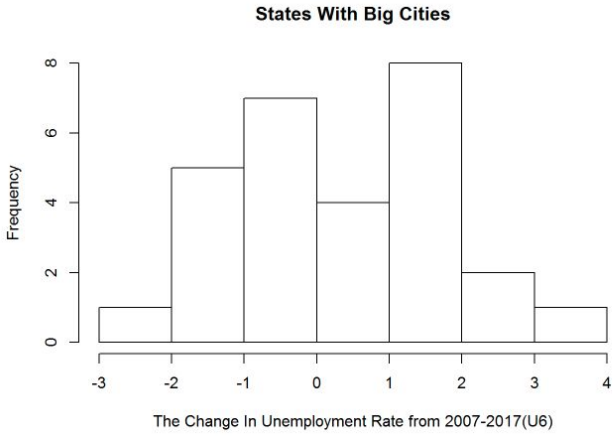
The other data set that I will use is one that I created using facts from the dataset found here(United States Census Bureau factfinder.census.gov). I call this dataset "citiesover500_data" in my supplemental materials. I made this dataset the hard way by manually inputting each state, whether they had a city with a population greater than 500,000, and how many cities with a population the state possesses. As stated in the last sentence, The columns in this dataset are the states, whether it has a city over 500,000 people(denoted Y or N for yes and no) and the number of cities in that state that meet that qualification. I created this to test the hypothesis that people migrated to highly populated metropolitan areas to find jobs, and those states with big cities recovered differently from the 2008 recession than states without these large cities.

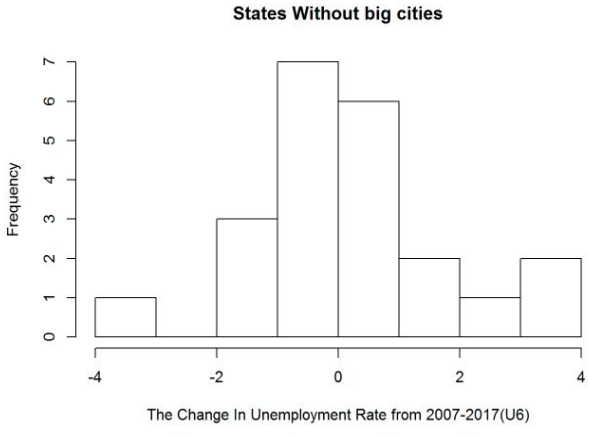
Note: any in-text citations denoted with a * point readers to more information on the method described

Method

The first action I took with the "myprojectdata" data was to clean up the unemployment data set in order to use it more efficiently. I selected the columns that had states and their corresponding unemployment(U6) rates. I took their 2007 unemployment(U6) rate and subtracted that from their 2017 unemployment(U6) rate to get change as a raw statistic that could be tested. I then merged this table with the one I created("citiesover500_data") that states whether each state has a city over 500,000 people in it, or doesn't.

Once I got past this point, I sorted by having states possessing a big city or not. I found that 22 states had a big city while 28 did not possess one. I calculated the means for both of these samples. I will want to see if there's a significant difference between the recovery(defined as the difference between 2007 and 2017 unemployment rates) in states that have greater than 500,000 population cities versus those without these cities. Both of these samples are pretty small, so I will compare the t-test and Wilcoxon-Mann-Whitney test in the simulations section to see which one will be better to use as a test for this hypothesis. The t-test has a few assumptions that could be troublesome for this analysis. One is that it must be randomly sampled, which we can assume because of the computational method of the t-test in Rmarkdown. Another one is that the samples are reasonably large which will present a problem as the samples are relatively small. 28 and 22 are relatively small sample sizes that barely satisfy the Central Limit Theorem, which states that if we sample from a population using a sufficiently large sample size, the mean of the samples (also known as the *sample population*) will be normally distributed (assuming true random sampling)(Nedrich spin.atomicobject.com)*. Finally, the last assumption follows from the CLT theorem, as the last assumption is that the underlying data follows a normal distribution. These are two histograms showing the change in U6 unemployment separated by states with big cities, and states without big cities.





Note: any in-text citations denoted with a * point readers to more information on the method described

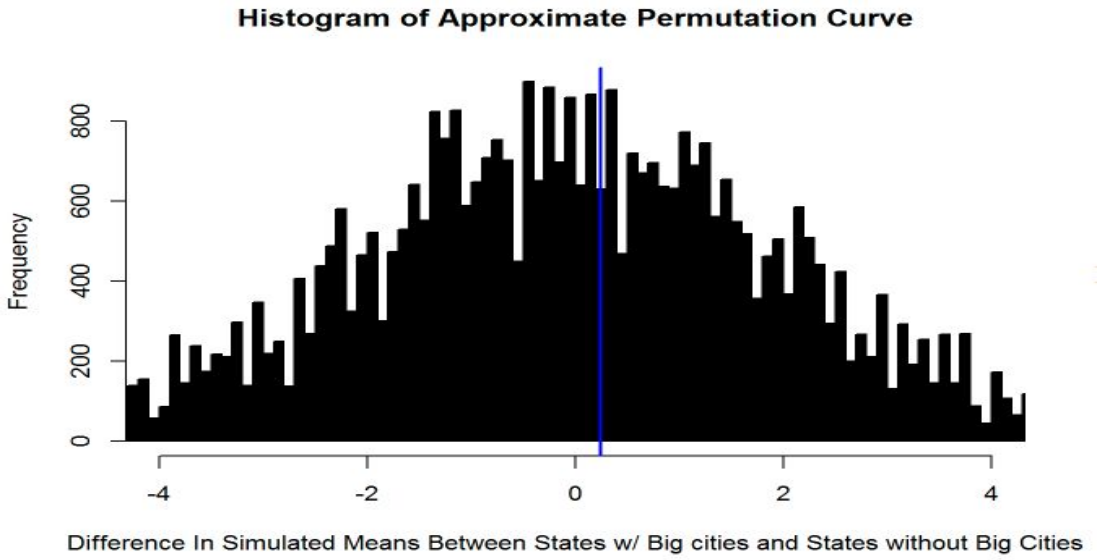
From this data, it would appear that the two samples follow a relatively similar pattern to a normal distribution. The only problem is it isn't that smooth all the way through, and state's with big cities have a big dip in the center, but the rest of the histogram seems to look approximately normal. The states without big cities seem to have a rather convincing normal distribution, but with small samples, we must be careful. It wouldn't be absurd to assume that with many more samples the relative change would follow a normal distribution. In fact, this is the rule of the Central Limit Theorem, which we discussed earlier.

The other test to see if the population means are significant is the Wilcoxon-Mann-Whitney(WMW) test. This is a two-sample rank test and can be used as long as the two samples are independent of each other. In this case, we will assume the states are independent of each other. The WMW test is a distribution-free permutation test, so it doesn't depend on the normal distribution like the t-test.

Simulations

The focus of the simulations section in this paper will be deciding which hypothesis test I will rely on for this research paper. I am comparing the t-test and Wilcoxon-Mann-Whitney test to see which one is more suitable for the dataset that I have chosen. The way I will be doing this is using the Monte Carlo approach for calculating the power for each test assuming the underlying distribution of the data is normal, and then that the underlying data follows a t-distribution.

The first thing I did to look at the distribution was permutation testing manually, by resampling the collected data myself, and randomly choosing whether they were going to be in the column with big cities or column without big cities. If there is no difference between the means, there shouldn't be any difference between the two columns. Then I took the difference of the means and repeated this process for 2000 iterations to get the approximate permutation curve of our distribution. The resulting graph I got was this:



Note: any in-text citations denoted with a * point readers to more information on the method described

Monte Carlo simulation performs random sampling and conducts a large number of experiments on a computer to help us see what distribution our inputs might follow("Chapter 8 Monte Carlo Simulation." web.mst.edu)*. The difference in the means when we use Monte Carlo methods above to create many random variables to help see the shape of our distribution appears to be approximately normal. In my many iterations of this test though, I have noticed that the two peaks next to the middle tend to stay rather large, suggesting there might be a deviation away from the normal distribution. The vertical blue line in this graph represents the difference in the sample means that we found originally for testing. We can see that the observed difference in sample means we found is right in the middle of this distribution. It does not appear to have a significant difference from the mean of this simulation.

To help decide between the WMW test and the t-test, we'll need to assess power under different distributions. Power is defined as the probability of rejecting the null hypothesis when the alternative hypothesis is true. We want this value to be high, and consistent through many types of distribution. Estimating power is one of the most essential tools for testing hypotheses(stats.idre.ucla.edu "Introduction to Power Analysis")*. We will use a Monte Carlo approach to approximate and simulate data using the sample means and standard deviations from our observed data under the normal distribution. We will set the significance level to .05 in this case. The significance level is also known as the probability of rejecting the null hypothesis when the null hypothesis is true. For this experiment, we are going to fix it to .05 in order to estimate the power of each test under different distributions. When we assume the underlying data we collected has a normal distribution, the t-test has a slightly higher power than the Wilcoxon-Mann-Whitney(WMW) test. But only by a very small amount, and sometimes the WMW test would come back with a higher power in simulations than the t-test. This can be seen and tested in my supplemental data.

I then used the same Monte Carlo estimation method, now assuming that the populations have an underlying t distribution. I gave both states without big cities and states with big cities their respectful degrees of freedom(27 and 22). Now I see the power differences between WMW and t-test for this t distribution. I notice that they again are close-ish together, but fluctuate a lot. The difference between the WMW results for both distributions tends to be less than the difference between the t-test results for both distributions, so I will pick the WMW-test for this experiment to determine whether there is a significant difference in mean U6 unemployment change from 2007-2017 between big city states and states without big cities.

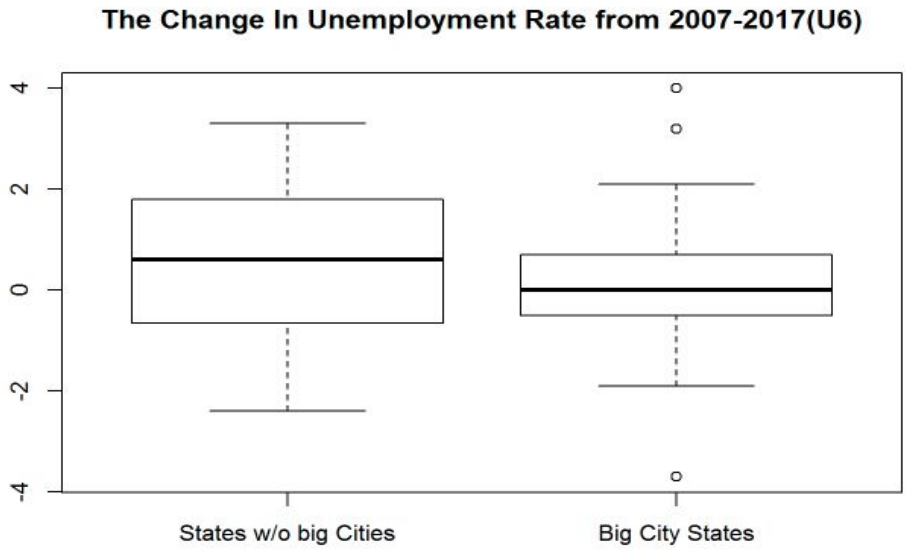
Note: any in-text citations denoted with a * point readers to more information on the method described

Analysis

From the simulations section, we have determined that we will be using the Wilcoxon-Mann-Whitney test for the research question of whether big city states have had a significantly different recovery(change in unemployment U6 from 2007 -2017) than states that do not possess big cities. The WMW test is a perfect fit as it doesn't require a significantly large sample like the t-test, or a normally distributed population. Although we have signs from the previous section that the underlying distribution of these samples might be normal, it is better to err on the side of caution. Especially when dealing with small sample sizes of 22 and 28, that could belong to many types of distribution.

When I sorted through the data and performed a WMW test on the change in unemployment(U6), I got a resulting p-value of .617. This is way above our stated rejection value in the simulation section of .05 and seems to imply that there is not a significant difference in change in unemployment(U6) between big city states and states with no big cities. The statistical decision from the WMW test would be that, at the 5% significance level, we fail to reject the null hypothesis that the population mean change in unemployment(U6) from 2007 to 2017 was the same for big city states and states without any big cities.

Before we make any absolute conclusions, it's essential to look at the spread of the data and determine if there's anything strange going with the data results we are looking at for this experiment. The boxplot to the right shows the spread of big city states and states without big cities.

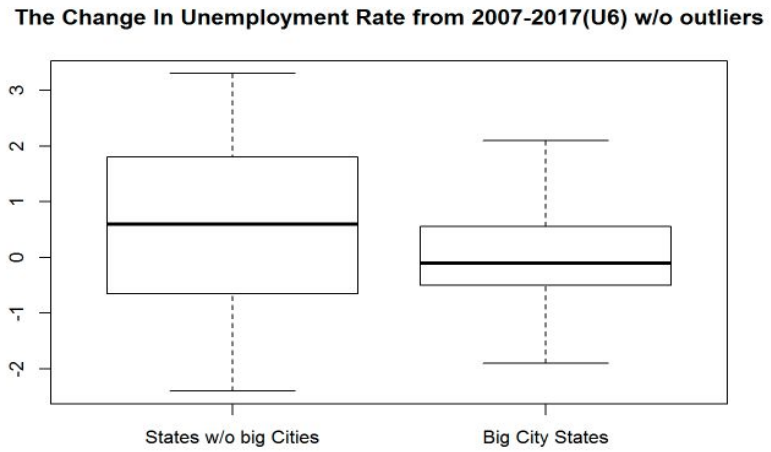


From this boxplot, we can notice a few important things. The boxplots look a bit different. States without big cities had a mean change of 0.4357143 in unemployment(U6) rate from 2007 to 2017, while states with big cities had a mean change of 0.1909091. This gives us a difference between these means of 0.2448052. With a small sample like this, it makes sense that

we would fail to reject they were significantly different given that the difference is such a low value. We can also see from this boxplot that states with big cities have a very tight boxplot, and three big outliers. We should investigate these outliers to see what states caused these outliers.

The outliers were New Mexico which gained 4 points in unemployment over this period, Nevada which gained 3.2 points of unemployment over this period, and the final outlier was the great state of Michigan which lost 3.7 points in unemployment over this period. After researching, I can't seem to find any empirical explanation for these substantial changes in New Mexico or Nevada. Michigan, on the other hand, had the second highest unemployment rate in 2007, so it makes sense that it would fall by a reasonable amount. The high unemployment rate was the result of many car manufacturers leaving and the GM bankruptcy that left many without work.

The question we need to address now is whether the test becomes any different when we filter out these outlier states. Michigan is a unique case, as they felt the effects of the recession much earlier. Now with removed outliers, we can repeat the process and see if there are any new conclusions we can make. This is the new boxplot with outliers removed.



Now the means are a little more differentiated. Big City States now have a mean change of 0.03684211, and now the difference of the means is 0.3988722. I will conduct another WMW test to see what the p-value looks like without these big outliers. The p-value determined by the WMW test in this case was .3976. This is still a relatively high p-value, but much smaller than the original p-value of .617. It gives a little credence to the idea that these big city states might have had a little bit better time recovering due to their highly populated urban centers, but these small sample sizes hurt us in making any conclusions about there being a significant difference. The statistical decision, in this case, would be again that, at the 5% significance level, we fail to

reject the null hypothesis that the population mean change in unemployment(U6) from 2007 to 2017 was the same for big city states and states without any big cities.

Now, that we have sorted outliers out, and seen that it makes a little difference but not enough to change the statistical decision. We can turn our attention to the number of big cities that individual states have, and see if that might have any impact on the unemployment rate. There are only three states with more than one city with a population greater than 500,000. They are shown below.

state	2007	2017	change	over500	numofcities
california	9.9	9.8	-0.1	Y	6
tennessee	8.0	7.9	-0.1	Y	2
texas	7.7	8.2	0.5	Y	6

Initially, I set out to include the number of cities estimate in my analysis, and maybe build a model where I could estimate how the unemployment rate would change based on the number of big cities a state has. I didn't expect for there to be no significant difference between the population mean change in unemployment rate between 2007 and 2017 for big city states and states without big cities. The difference appears to be even less significant with states that have multiple big cities vs. those that have one or zero big cities. When I apply the WMW test to these two datasets, I obtained a p-value of .9834 which would not be a significant p-value for any reasonable rejection value of alpha. A sample size of only three states is very hard to draw a conclusion from, and not a significant enough sample size to validate testing. On the other hand, these states did have relatively small changes in the unemployment rate, so it might beg the question to future researchers whether these states with multiple cities can have a significant difference on the recovery of an economy/unemployment.

Discussion

The main conclusion that my research project has come to is that there is not enough evidence to reject the null hypothesis that states with big cities and states without big cities had the same population mean change in recovery from 2007 to 2017 in terms of unemployment rate(U6). We went through the data I collected, simulations I used to determine which test was going to be more effective for the samples I was testing, and finally analyzed what this all meant.

The data I used was from the Bureau of Labor Statistics, in which they reported the annual unemployment rate for each state. I used the U3 and U6 measurements. Then I evaluated the power of the WMW test and t-test to see which of these would be a better test for finding out if the mean change in unemployment was different in big city states vs. states with no big cities.

Note: any in-text citations denoted with a * point readers to more information on the method described

As stated above I found that there was no evidence to suggest a significant difference for the mean unemployment rate in states with big cities, and states without big cities.

This research paper did not find that having big cities had an impact on lowering or raising the unemployment rate, but it did end up bringing some other issues to investigate. The sample means were reasonably different, and when we took out outliers, they become even more different. There might be something else going on with these states that I did not fully capture. One idea would be comparing the unemployment rate to population growth in each state, or maybe Gross Domestic Product(GDP) growth in each state. There would need to be a collection of the GDP values for each state in order to test that hypothesis.

One flaw I realized in my project is that although unemployment is a standard indicator of economic health. It is possible to have a low unemployment rate and a weak economy. States like Mississippi and Alabama suffer from an issue referred to as "Brain Drain". The idea is that highly qualified workers leave their economy to go to states with more like-minded and skilled people. These people might even prefer going to these states, even without a job offer. One example of a reason for leaving the state would be education. States like Mississippi might have low unemployment rate not because they have an excellent economy, but because their population is so small that all the jobs are being occupied. There aren't enough highly qualified applicants, so high-level jobs are being filled by people that are underqualified for the position. This is where a GDP measure might help to flesh out the economic ideas that I set out to answer. We could see if the actual economy is growing, not just the unemployment rate.

It might also be useful to look at city population growth in each state, and whether that affected the economy. My research project can be a jumping point for many other people to test and see what differences there are among unemployment rates among states. It can start a conversation about how states recovered from the recession. A future project could use a more broad scope and utilize more data such as GDP per state and population shifts to make different inferences about how states recovered from the recession. My hope is that this research project was able to help inform and create a conversation about how states recovered from the recession.

Note: any in-text citations denoted with a * point readers to more information on the method described

Works Cited

Author: https://www.r-bloggers.com/author/is-r/. "US State Maps Using map_data()." *R-Bloggers*, 11 Dec. 2012, www.r-bloggers.com/us-state-maps-using-map_data/.

*"Chapter 8 Monte Carlo Simulation." *Probabilistic Engineering Design*, Missouri University of Science and Technology, 2 Feb. 2017, web.mst.edu/~dux/repository/me360/ch8.pdf.

Horsley, Scott. "Ahead Of Trump's First Jobs Report, A Look At His Remarks On The Numbers." *NPR*, NPR, 29 Jan. 2017, www.npr.org/2017/01/29/511493685/ahead-of-trumps-first-jobs-report-a-look-at-his-remarks-on-the-numbers.

*"INTRODUCTION TO POWER ANALYSIS." *UCLA Institute for Digital Research & Education*, stats.idre.ucla.edu/other/mult-pkg/seminars/intro-power/.

Leeper, Thomas J. "Permutation Tests." *Thomas J. Leeper*, thomasleeper.com/Rcourse/Tutorials/permutationtests.html.

*Nedrich, Matt. "An Introduction to the Central Limit Theorem." *Atomic Spin*, 30 Sept. 2016, spin.atomicobject.com/2015/02/12/central-limit-theorem-intro/.

Rafter, Dan. "The Numbers Show It: Jobs, People Flocking to Big Cities." *RE Journals*, REjournals, 20 July 2018,

Note: any in-text citations denoted with a * point readers to more information on the method described

www.rejournals.com/the-numbers-show-it-jobs,-people-flocking-to-big-cities-2080720.

United States Bureau of Labor Statistics. *Bureau of Labor Statistics*, 26 Apr. 2019, www.bls.gov/lau/stalt_moave.xlsx .

United States Census Bureau. "Annual Estimates of the Resident Population for Incorporated Places of 50,000 or More, Ranked by July 1, 2017 Population: April 1, 2010 to July 1, 2017 - United States -- Places of 50,000+ Population." *United States Census Bureau* , May 2018, factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?src=bkmk.